W230–W232 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki412

DiANNA: a web server for disulfide connectivity prediction

F. Ferrè[1] and P. Clote[1,2,*]

[1]Department of Biology and [2]Department of Computer Science (courtesy appointment),
Boston College, Chestnut Hill, MA 02467, USA

Received February 13, 2005; Revised and Accepted March 21, 2005

ABSTRACT

Correctly predicting the disulfide bond topology in a protein is of crucial importance for the understanding of protein function and can be of great help for tertiary prediction methods. The web server http://clavius.bc.edu/~clotelab/DiANNA/ outputs the disulfide connectivity prediction given input of a protein sequence. The following procedure is performed. First, PSIPRED is run to predict the protein's secondary structure, then PSIBLAST is run against the non-redundant SwissProt to obtain a multiple alignment of the input sequence. The predicted secondary structure and the profile arising from this alignment are used in the training phase of our neural network. Next, cysteine oxidation state is predicted, then each pair of cysteines in the protein sequence is assigned a likelihood of forming a disulfide bond—this is performed by means of a novel architecture (diresidue neural network). Finally, Rothberg's implementation of Gabow's maximum weighted matching algorithm is applied to diresidue neural network scores in order to produce the final connectivity prediction. Our novel neural network-based approach achieves results that are comparable and in some cases better than the current state-of-the-art methods.

INTRODUCTION

Disulfide bonds are covalently bonded sulfur atoms from nonadjacent cysteine residues, which stabilize the protein structure and are often found in extracytoplasmatic proteins. The knowledge of cysteine connectivity (i.e. which, if any, pairs of cysteines form a bond in a given protein sequence) can reduce greatly the conformational space for protein structure prediction algorithms. Moreover, as shown by Chuang and co-workers (1), a similar disulfide connectivity pattern frequently implies a structural similarity even when the sequence similarity is undetectable. Notwithstanding, only a few attempts have been made to solve this problem. In contrast, many methods have been developed for the related, but simpler problem of cysteine oxidation state prediction, i.e. to determine the cysteines that are involved in a disulfide bond, without predicting the connectivity pattern. Recent methods based on machine learning techniques have reached an outstanding accuracy of 90% on certain test data (2–5). In spite of this, accuracy for the disulfide connectivity problem remains measured. The reason for this is simple—amino acids that flank half-cystines (disulfide-bonded cysteines) are quite different from those that flank free cysteines (non-bonded cysteines) (6,7). In contrast, the residues that flank two incorrectly paired half-cystines are quite similar to those that flank the half-cystines in a disulfide bond. Two recent and remarkable papers based on different approaches (8,9) outperform early attempts by Fariselli and co-workers (10,11). The Vullo and Frasconi method (9) uses recursive neural networks (12) to score undirected graphs that represent cysteine connectivity. The method of Zhao and co-workers (8) is based on recurrent patterns of sequence separation between bonded half-cystines. Web servers that allow online disulfide connectivity prediction are available for Vullo/Frasconi (http://cassandra.dsi.unifi.it/cysteines) and, as a prototype, for Fariselli/Casadio (http://gpcr.biocomp.unibo.it/cgi/predictors/cys-cys/pred_dconcgi.cgi). Here, we describe a web server for disulfide connectivity prediction that implements our novel approach, which results in comparable and sometimes better than the state-of-the-art methods (8,9). Algorithm details and performance of the method are described previously by Ferrè and Clote (13).

METHODS

The stand-alone program for disulfide connectivity prediction, implemented in our web server DiANNA (for DiAminoacid Neural Network Application), uses a three-step procedure. First, a neural network is trained to recognize cysteines in an oxidized state (sulfur covalently bonded) as distinct from cysteines in a reduced state (sulfur occurring in reactive

*To whom correspondence should be addressed. Tel: +1 617 552 1332; Fax: +1 617 552 2011; Email: clote@bc.edu sulfhydryl group SH), based on the previous work by Fariselli *et al*. (14) only those monomers that have at least two predicted half-cysteines are submitted to the second step. The neural network input is a window of size w centered at each cysteine in the sequence. This first filtering step is called Module A. Then, a second neural network (Module B) is used to score each pair of symmetric windows of size w, each one centered at a cysteine in the input sequence. The network input contains evolutionary information, i.e. each residue is encoded by 20 input units corresponding to the PSIBLAST-computed profile row (obtained from the multiple alignment of the input sequence against the non-redundant SwissProt), and secondary structure information, computed using PSIPRED (15) and encoded in unary format by the addition of three input units, e.g. helix is encoded 1 0 0, coil is 0 1 0 and sheet is 0 0 1). Using secondary structure information leads to a marked improvement and is justified by the bias in the secondary structure preference of free cysteines and half-cystines (16). The architecture of the Module B neural network is as follows. Given an encoded input containing secondary structure information, thus having $w \times 23$ input units, we designed a first hidden layer containing $\binom{w}{2} = w(w-1)/2$ units, one for each pair $1 \leqslant i < j \leqslant w$ of positions, with connections to input units representing the profile for residues at position i, j and secondary structures at those positions. Thus, each of the $w\,(w-1)/2$ hidden units in the first hidden layer (the diresidue layer) is connected to 2 (20 + 3) = 46 input units (Figure 1). A second hidden layer, containing five units, all fully connected with those of the first hidden layer, is then fully connected to the single output unit. We designed this unusual neural network architecture, with the aim of emphasizing the signal that arises when using diresidue position-specific scoring matrices (13), i.e. for all windows of length w, for positions $1 \leqslant i < j \leqslant w$ and amino acids a, b, we consider the frequency of occurrence of amino acid a in position i when amino acid b is found in position j; moreover, though there are many hidden units, the training phase is still reasonably fast since the diresidue layer is not fully connected with the input layer.

Finally, following Fariselli and Casadio (10), our algorithm applies the Edmonds–Gabow maximum weight matching algorithm (17,18), using Ed Rothberg's implementation wmatch (http://elib.zib.de/pub/Packages/mathprog/matching/weighted), to the weighted complete graph, whose nodes are half-cystines and whose weights are values output from the neural network of Module B. This last step is called Module C.

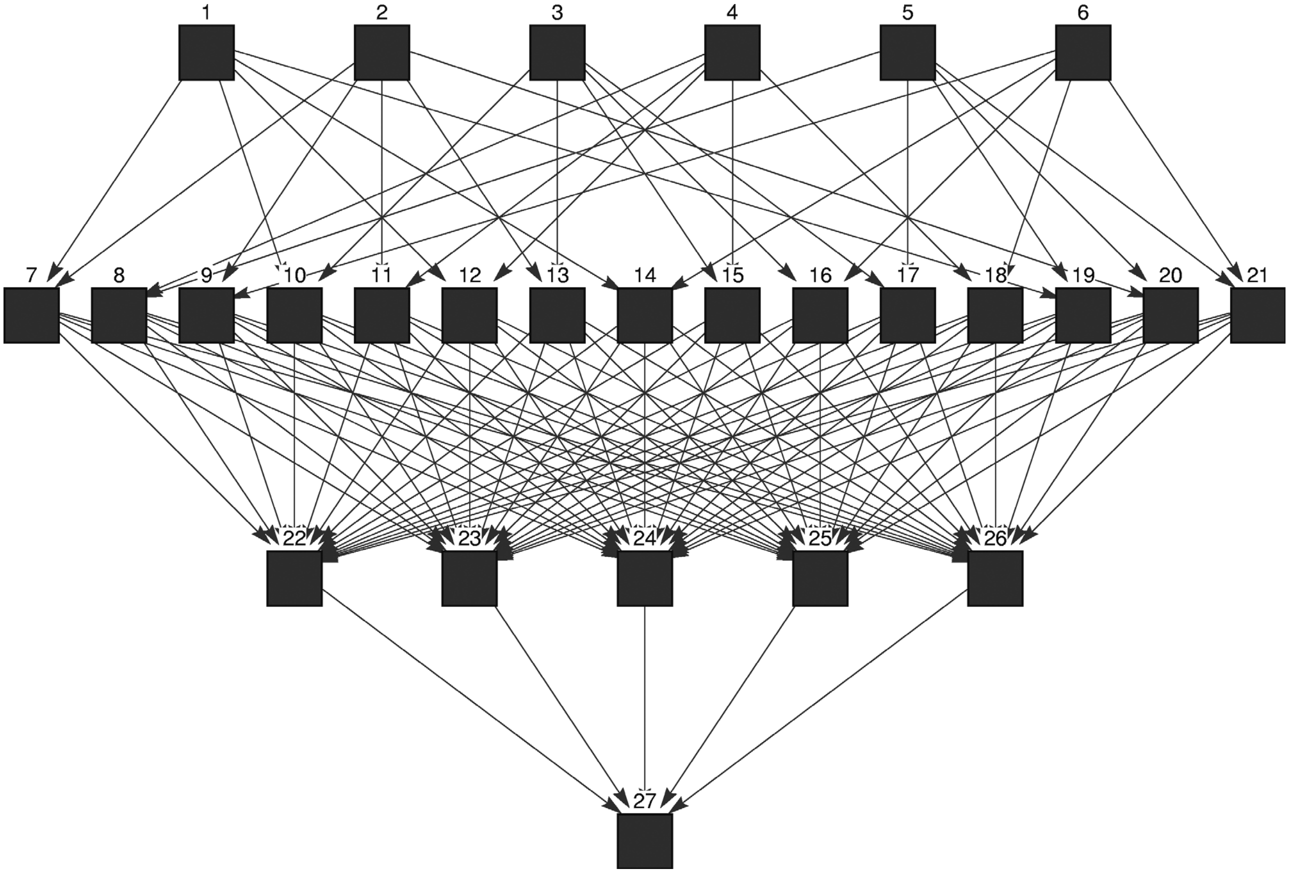
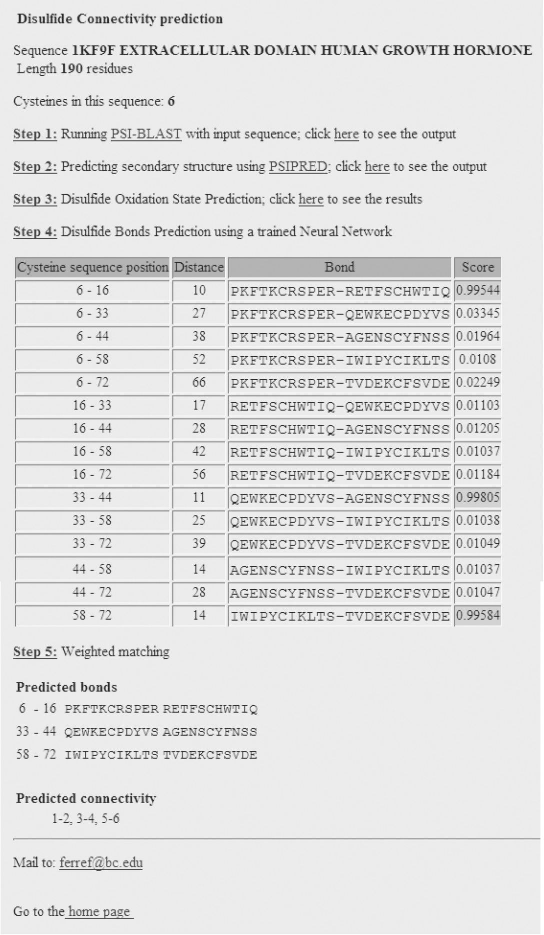


Figure 1. A toy example of the diresidue neural network architecture. Six input units (named 1, . . , 6) are connected to the $\binom{w}{2}$ units of the first hidden layer (7, . . , 21), called the diresidue layer. Each pair of input units is connected to a distinct unit in the diresidue layer. The units of the diresidue layer are then fully connected to the five units (22, . . , 26) of the second hidden layer, which are fully connected to the single output unit. Using the second hidden layer provided a better performance than connecting the diresidue layer units directly to the output unit. In the DIANNA application, each residue is encoded by 23 input units (20 encoding the evolutionary information and 3 for the secondary structure information); therefore, each unit in the diresidue layer is connected to 23 + 23 = 46 input units that code a pair of residues.



Disulfide Connectivity prediction

Sequence **1KF9F EXTRACELLULAR DOMAIN HUMAN GROWTH HORMONE**
Length **190** residues

Cysteines in this sequence: **6**

Step 1: Running PSI-BLAST with input sequence; click here to see the output

Step 2: Predicting secondary structure using PSIPRED; click here to see the output

Step 3: Disulfide Oxidation State Prediction; click here to see the results

Step 4: Disulfide Bonds Prediction using a trained Neural Network

Cysteine sequence position	Distance	Bond	Score
6 - 16	10	PKFTKCRSPER-RETFSCHWTIQ	0.99544
6 - 33	27	PKFTKCRSPER-QEWKECPDYVS	0.03345
6 - 44	38	PKFTKCRSPER-AGENSCYFNSS	0.01964
6 - 58	52	PKFTKCRSPER-IWIPYCIKLTS	0.0108
6 - 72	66	PKFTKCRSPER-TVDEKCFSVDE	0.02249
16 - 33	17	RETFSCHWTIQ-QEWKECPDYVS	0.01103
16 - 44	28	RETFSCHWTIQ-AGENSCYFNSS	0.01205
16 - 58	42	RETFSCHWTIQ-IWIPYCIKLTS	0.01037
16 - 72	56	RETFSCHWTIQ-TVDEKCFSVDE	0.01184
33 - 44	11	QEWKECPDYVS-AGENSCYFNSS	0.99805
33 - 58	25	QEWKECPDYVS-IWIPYCIKLTS	0.01038
33 - 72	39	QEWKECPDYVS-TVDEKCFSVDE	0.01049
44 - 58	14	AGENSCYFNSS-IWIPYCIKLTS	0.01037
44 - 72	28	AGENSCYFNSS-TVDEKCFSVDE	0.01047
58 - 72	14	IWIPYCIKLTS-TVDEKCFSVDE	0.99584

Step 5: Weighted matching

Predicted bonds

```
6  - 16  PKFTKCRSPER RETFSCHWTIQ
33 - 44  QEWKECPDYVS AGENSCYFNSS
58 - 72  IWIPYCIKLTS TVDEKCFSVDE
```

Predicted connectivity

1-2, 3-4, 5-6

Mail to: ferref@bc.edu

Go to the home page

Figure 2. Output from DIANNA when given as input the sequence for human growth hormone receptor (SwissProt ID GHR_HUMAN, PDB code 1kf9 chain F). This protein has 6 cysteines that form 3 disulfide bonds, with connectivity pattern 1–2, 3–4, 5–6 (between cysteines 6 and 16, 33 and 44, 58 and 72). The upper portion of the output page reports the Module B score (see text) for each pair of cysteines, ranging from 0 to 1 (scores >0.9 are highlighted). In the lower portion, the proposed connectivity (i.e. the Module C output) is shown.

SERVER DESCRIPTION

The web server takes as input a protein sequence in FASTA format and can output the following: (i) oxidation state prediction for all the cysteines in the input sequence, using our implementation of the neural network described in (14) (Module A); (ii) a score for each pair of cysteines in the input, obtained by our diresidue neural network (Module B); (iii) the disulfide connectivity prediction obtained using the maximum weighted matching algorithm (Module C) applied to the scores of Module B. The user is warned if Module A predicts less than two half-cystines in the input sequence. A statistical evaluation of the connectivity prediction is not attempted. A sample output is shown in Figure 2.

DISCUSSION

Trained and tested on a list of proteins having at most five and at lest two bonds, equivalent to those used in (9,11), the software achieves a rate Q_p of 49% for perfect predictions (i.e. the fraction of proteins for which there are no false-positive or false-negative predictions made), 86% accuracy and 51% Matthews' correlation coefficient (13). For proteins having two and four bonds, the fraction of perfect predictions improves to 62 and 55%, respectively. Although future improvement for disulfide connectivity is still desired, our approach is nonetheless reliable when used on proteins having a relatively small number of disulfide bonds.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by Boston College.

Conflict of interest statement. None declared.

REFERENCES

1. Chuang,C.C., Chen,C.Y., Yang,J.M., Lyu,P.C. and Hwang,J.K. (2003) Relationship between protein structures and disulfide-bonding patterns. *Proteins*, **53**, 1–5.
2. Martelli,P.L., Fariselli,P. and Casadio,R. (2004) Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. *Proteomics*, **4**, 1665–1671.
3. Martelli,P.L., Fariselli,P., Malaguti,L. and Casadio,R. (2002) Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. *Protein Eng*., **15**, 951–953.
4. Mucchielli-Giorgi,M.H., Hazout,S. and Tuffery,P. (2002) Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins*, **46**, 243–249.
5. Chen,Y.C., Lin,Y.S., Lin,C.J. and Hwang,J.K. (2004) Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. *Proteins*, **55**, 1036–1042.
6. Muskal,S.M., Holbrook,S.R. and Kim,S.H. (1990) Prediction of the disulfide-bonding state of cysteine in proteins. *Protein Eng*., **3**, 667–672.
7. Fiser,A., Cserzo,M., Tudos,E. and Simon,I. (1992) Different sequence environments of cysteines and half cystines in proteins. Application to predict disulfide forming residues. *FEBS Lett*., **302**, 117–120.
8. Zhao,E., Liu,H.L., Tsai,C.H., Tsai,H.K., Chan,C.H. and Kao,C.Y. (2004) Cysteine separations profiles (CSP) on protein sequences infer disulfide connectivity. *Bioinformatics*, **20**, 653–659.
9. Vullo,A. and Frasconi,P. (2004) Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics*, **20**, 653–659.
10. Fariselli,P. and Casadio,R. (2001) Prediction of disulfide connectivity in proteins. *Bioinformatics*, **17**, 957–964.
11. Fariselli,P., Martelli,P.L. and Casadio,R. (2002) A neural network based method for predicting the disulfide connectivity in proteins. In Damiani,E. *et al*. (eds), *Knowledge Based Intelligent Information Engineering Systems and Allied Technologies (KES)*. IOS Press, Amsterdam, pp. 464–468.
12. Frasconi,P., Gori,M. and Sperduti,A. (1998) A general framework for adaptive processing of data structures. *IEEE Trans. Neural Netw*., **9**, 768–786.
13. Ferrè,F. and Clote,P. (2005) Disulfide connectivity prediction using secondary structure information and diresidue frequencies. *Bioinformatics*, in press.
14. Fariselli,P., Riccobelli,P. and Casadio,R. (1999) Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. *Proteins*, **36**, 340–346.
15. Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol*., **292**, 195–202.
16. Petersen,M.T., Jonson,P.H. and Petersen,S.B. (1999) Amino acid neighbours and detailed conformational analysis of cysteines in proteins. *Protein Eng*., **12**, 535–548.
17. Gabow,H. (1973) *Implementation of Algorithms for Maximum Matching on Nonbipartite Graphs*. PhD Thesis, Stanford University, CA.
18. Lovasz,L. and Plummer,M. (1985) *Matching Theory. B.V. North Holland Mathematical Studies*. Vol. 121, Elsevier Science Publishers.